ips
Italian Political Science, VOLUME 12, ISSUE 1, JUNE 2017
SISP SOCIETÀ ITALIANA DI SCIENZA POLITICA

Two Decades of Political Science Research Assessment: the Dutch Experience

Rudy B. Andeweg
LEIDEN UNIVERSITY

A High Level of Acceptance

Today, there is a general acceptance of, or at least resignation about, Dutch Research Assessments, whether in political science or in other disciplines. Research assessment exercises started in the Netherlands in 1993, and are held every six years. To a large extent, research assessments are a non-issue. In comparison to the experience of political scientists in many other countries, this may seem surprising, but several factors help account for this counter-intuitively high level of satisfaction.

The single most important factor underlying the acceptance undoubtedly is the simple fact that the role of the government in organizing, administering and supervising the assessments is marginal. A recent report by an independent think tank concluded that nowhere in Europe is the involvement of the government or other state actors as minimal as it is in the Netherlands (Van Drooge et al. 2013). The universities alone are responsible for the assessments. The *Standard Evaluation Protocol* (SEP), which outlines the aims and procedures of the research assessments, has been developed by the Dutch Association of Universities (VSNU) together with the Dutch Science Foundation (NWO) and the Royal Academy of Sciences (KNAW), organizations that are beyond the direct control of the government. The introduction to the latest edition of the Standard Evaluation Protocol (2014) mentions that it was presented to the Minister of Education, but merely out of politeness. Neither the Minister nor her civil servants had been involved in setting the evaluation criteria, and even the obligation to send a copy of each completed assessment exercise to the Education Minister has been dropped several years ago.

The universities define the research units that are to be subjected to an assessment exercise; each university decides whether its research units will be assessed in a stand-alone exercise, or whether they will be part of a nation-wide comparative assessment of research in that particular discipline. The most recent Political Science Research Review (Verdun et al 2014), for example, did not include the Department of Political Science at Radboud University Nijmegen, because that university had opted for a stand-alone assessment of its political science research programme. The universities decide on the composition of the peer review committee that will conduct the assessment, as long as it is an international committee and its members have no conflict of interest with any of the departments, and often the university executives will delegate the search for committee

Volume 12, Issue 1, pp. 29–35.

members to representatives of the departments concerned. The universities also provide logistic and administrative support to the assessment committees, and through the Dutch Association of Universities they have set up an independent agency QANU (Quality Assessment of Netherlands' Universities) which specializes in offering such support. It is fair to conclude that the Dutch Research Assessments are free from governmental interference.

A second reason for the general acceptance of the research assessments is that they hardly have any direct consequences for the scholars whose work is evaluated. To some extent this is related to the lack of government interference. The government could still use the reports, which are made public, to shape its funding decisions, but it does not. Even the universities do not attach direct consequences to the assessment outcomes. Doing so would contravene the twin aims of the assessment exercises: accountability for the use of taxpayer money, and improvement of the research units involved. These aims are explicitly stated by the universities themselves, which limits their ability to punish a research unit for poor assessment results by reducing funding or closing down departments. The only direct consequence that I have been able to find is for the accreditation of Research Master Programmes. In the Netherlands, Master programmes in all but a few disciplines are one year programmes. Ministerial permission is required for the start of a two-year Research Master catering to selected talented students, primarily potential PhD candidates. The Minister bases such decisions on the recommendation of (re-)accreditation panels, and one of the criteria used is having obtained high scores in the most recent research assessment exercise.

There are more indirect consequences. Departments take the research assessments very seriously because they affect their reputation. Getting a bad evaluation, or even a good evaluation that is significantly below the evaluations of other departments in the same discipline, has a negative effect on the department's reputation, which is feared to weaken a department's potential to recruit good PhD candidates and faculty, and to weaken its potential to receive research grants from the science foundation. Still, it would seem that the absence of direct sanctions helps explain the relative satisfaction.

In the Netherlands, there is a parallel scheme for the assessment of teaching quality, and there seems to be more concern about the nature and aims of those reviews. In any given six-year cycle, most departments will be evaluated twice, once on the quality of their research, and once on the quality of their teaching. Although the teaching quality assessments are also organized by the universities themselves, the reports are used by the Minister of Education and her Inspectorate. In 1994-1995, such an assessment report was used by the Inspectorate and the Minister to threaten to withdraw the accreditation of the Bachelor programme in political science at Radboud University Nijmegen – a threat that was lifted only after the University promised major reforms. Moreover, the outcomes of the teaching quality assessments are used by others, including commercial publishers, who draw up rankings of Bachelor programmes to aid prospective students in choosing which university to go to. As the funding of universities, and of departments within universities, is largely determined by student numbers, a poor teaching quality assessment may have immediate effects on the intake of students, and thus on the funding, of departments. So the immediate consequences of the teaching quality assessments are much more important than those of the research quality assessments.

Criticisms

The fact that research assessments are hardly controversial in Dutch academia does not mean that there are no criticisms of aspects of the assessment exercises. Some of the criticisms have led to adaptations in the regularly updated Standard Evaluation Protocol, but on others the process has been less responsive.

Administrative burden

A major complaint refers to the administrative burden. For each assessment, a department has to hand in a self-evaluation report. Such a report should contain quantitative information on the research input and output, conforming to very specific standardized criteria. Occasionally this requires collecting new data or transforming existing data to meet the Standard Evaluation Protocol's criteria – for example when a university employs different definitions of peer-reviewed/non-peer-reviewed publications, or national/international publications for its internal use. In addition, the self-evaluation report should contain a qualitative reflection by the department of its own research policy, publication strategy, etc. This should be presented in the form of a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis, and often prompts lengthy deliberation on finding the right balance between being honest and being strategic: being very honest makes it easy for the assessment committee to expose the department's weaknesses; being too strategic may prompt the committee to distrust the self-evaluation and to dig deeper itself.

An often used strategy to deal with this dilemma is to be quite honest about the department's weaknesses, but to start reforms to address these weaknesses just before the committee arrives for its site visit. The site visit itself is disruptive for a department, but it is brief. More work awaits the department after the assessment report has been published, as most university administrations will request a follow-up report from the department to show what will be done with the committee's recommendations. Moreover, most universities fear the effects of a negative assessment on their reputation, and require departments to organize a midterm assessment themselves in order to be able to address any vulnerabilities before the real assessment takes place. Although this is less burdensome than the real research assessment, it still requires compiling a self-evaluation document and discussing it with an external assessor, usually a trusted colleague from a university outside of the country.

Given the fact that assessments of research and teaching quality follow quite similar procedures, most departments have to write two self-evaluations, two follow up reports, and organize two midterm assessments in any given six-year cycle. Nothing has been done to alleviate this administrative burden.

The Improvement/Accountability Dilemma

As mentioned above, the stated aims of the Dutch research assessments are accountability and improvement. These aims are not contested, but in practice they are difficult to reconcile. In terms of accountability it is necessary that the assessment reports are given wide publicity, and include the evaluations of all research units in a given discipline. This makes it easy for the interested taxpayers to see what was done with their money. But such public and comparative reports may lead to posturing by departments rather than to frank SWOT analyses in their self-evaluation reports. Such reticent self-evaluations will hamper

assessment committees in identifying weaknesses and developing useful suggestions for improvement. Since 2003, universities are allowed to organize stand-alone research assessments, and an example of such an assessment exercise in political science was mentioned above. Even if such non-comparative assessments are made public, they do not attract the same amount of attention that the comparative reports attract. It could well be argued that stand-alone reports are preferable in terms of searching for improvement of research quality as there is less need for a department to act strategically. However, this comes at a cost in terms of accountability. Moreover, withdrawing from the national and comparative research assessment exercise is generally interpreted as an admission of weakness by one's colleagues. Nevertheless, the number of stand-alone research assessments has increased considerably. Across all disciplines 222 research assessments have taken place between 1994 and 2012, 136 of which were confined to just one university or research unit (Van Drooge 2013: 7). In political science, with the exception mentioned, comparative assessment exercises are still the norm.

One size fits all?

Originally, the assessment protocols made no allowance for differences between disciplines. The assessment criteria were largely based on what was customary in the technical and natural sciences. Research assessments were not alone in having this bias towards a publication culture that favours journal articles over books, English-language over Dutch-language publications, and multi-authored over single-authored publications. This bias has had a marked impact on the publication culture within political science. Gradually, however, the protocols allow for greater variety and fine-tuning to the needs of the discipline being evaluated. In the most recent political science research assessment, for example, it was decided to use bibliometric data from Google Scholar rather than Web of Science, as the first has a better coverage of political science publications than the latter.

A recent report of the Royal Netherlands Academy of Arts and Sciences advocates to find a balance between uniform assessment criteria and taking into account the variety within the social sciences, by adopting a simple 2×3 table of assessment categories, and leaving it to each discipline to fill those categories with indicators that are relevant to that discipline (Bensing et al. 2013).

		Quality domains	
		Scientific quality	*Societal relevance*
Assessment Dimensions	*Output*	Output regarded by peers as of outstanding quality	Output for external target groups
	Utilization	Utilization by peers of researcher's output	Utilization by external target groups
	Recognition	Recognition by peers Researcher's output	Recognition by external target groups

It is too early to say whether this recommendation will be implemented and assessment criteria will be furthered tailored to the publication culture and the specific needs of political science and the other social sciences.

The Problem of Proxies

Research quality is a largely subjective concept for which no clear and generally accepted indicators are available. As a consequence, all indicators that are used in assessment exercises are proxies, and usually proxies of a quantitative nature: the number of publications, citation scores, the amount of external research funding, etc. There is increasing dissatisfaction with such quantitative criteria that almost by definition imply that 'more is better'. The concern is that it will lead to strategic behavior: mutually adding colleagues as coauthors so that all members of the department have more publications. In at least one Dutch political science department it has become the rule that the PhD supervisor is automatically listed as a coauthor of all publications of the PhD candidate. This led the most recent Assessment committee to conclude that 'there are questions for each of these Institutes about whether PhD candidates in their Programmes should publish together with their supervisors (and if so whether those publications should form part of their dissertation work)' (Verdun et al. 2013: 13).

Here too there has been some responsiveness to those concerns. Research units are asked to list what it considers its five best publications over the past six years, and assessment committees are expected to read them, although it is not always clear from the report that the committee actually did so. Of the four quality indicators used so far: (scientific quality, scientific productivity, societal relevance, and viability), the most quantitative indicator – productivity – has been dropped, and research integrity has been added.

Outcome inflation?

Although the research units that are assessed do not complain, it is perceived by policy-makers as a problem that the average scores that are used to summarize a department's research quality have gone up over the years, leaving very little variation between the research units that have been assessed. So far, the scores have been expressed on a scale of 1 to 5. On the indicator of quality, for example, the average score went up from 3.65 in the first assessment cycle in the 1990s to 4.39 in the most recent 2009-2015 cycle (Van Drooge 2013: 10). Cynics might surmise that this increase is correlated to the increase in stand-alone assessments, but a comparison between the average scores used in comparative and in stand-alone assessments shows that this is not the case.

In the most recent Political Science research assessment (Verdun et al. 2014) the variation in scores across departments is indeed small:

	Quality	Productivity	Relevance	Viability
Leiden University	4.5	5	4	4.5
Amsterdam University	4.5	4	4	4
Free University Amsterdam	4.5	4.5	4.5	4.5
Twente University	4	4	5	3.5

In response to what is perceived as 'score inflation', the scale has been redefined several times. From 1=poor, 2=unsatisfactory, 3=average, 4=good and 5=excellent, to 1=unsatisfactory, 2=satisfactory, 3=good, 4= very good and 5=excellent. In the next round the scale will be reversed and range from 4=unsatisfactory, 3=good, 2=very good, to 1=

world leading. It is hoped that such changes will also produce more variation in the scores awarded to various research units.

However, it is not clear whether the higher and more homogeneous scores indeed reflect score inflation. After all, it is one of the explicit aims of the research assessments to help improve the research quality at Dutch universities. If, after over twenty years of research quality assessments, quality would not have improved, this would not reflect well on the utility of the whole exercise. Similarly, as the room for improvement was greater for departments that started out with relatively low scores, it should not come as a surprise that there is less variation two decades later.

Impact

As they hardly have any direct consequences, it is not possible to measure the impact of the research assessments. Moreover, the introduction of research assessments in the early 1990s was but one element in the general professionalization of political science in the Netherlands. This professionalization was not only imposed from above by research assessments, by reducing the income that universities receive from the state directly, making them more independent on the competition for external research funding, etc., but it has also been initiated from below, by political scientists who sought to maintain or strengthen their reputation in an increasingly international environment. A recent overview of the development of Dutch political science is entitled 'from politicization to professionalization' (Andeweg & Vis 2015), and describes how professionalization has also been a reaction to political scientists growing tired of the ideological conflicts that plagued some of their departments (the two universities in Amsterdam and Nijmegen university in particular) from the 1960s to the 1980s. In that light, the undoubtedly positive outcome of professionalization and internationalization can only in part be attributed to the research assessments.

The other side of the coin is that the downside of professionalization and internationalization can also be blamed only partially on the research assessments. One of these negative side effects is the shift in the publication culture towards co-authored English-language articles in peer-reviewed journals. There are no intrinsic reasons for this shift from books to journals and for the increase in the average number of coauthors. It has less to do with increasing quality than with succumbing to the temptation to measure research quality by readily available bibliometric indicators. We have allowed ourselves to be taken hostage by a commercial firm: Thomson Reuters and its Social Science Citation Index!

The trend to publish more internationally, i.e. in English, does not have only negative consequences. After all, an English language publication is accessible to a much wider readership than a publication in Dutch, which brings a higher level of scrutiny and debate. This can only have beneficial consequences in terms of research quality. However, the shift in publishing from Dutch to English, and the higher threshold to readers because of the more sophisticated methodology used, has also meant that political science plays a significantly less prominent role in public debate in the Netherlands: science for science, rather than science for society. In the media, we see that historians and constitutional lawyers increasingly replace political scientists when journalists need expertise to explain current events.

The changes that already have been made to the Standard Evaluation Protocol, and the further changes that have been advocated, can be seen as efforts to address the negative effects of professionalization and internationalization: less emphasis on productivity and more attention to research integrity may help stop some of the strategic publishing choices that have emerged, and more attention to societal relevance may induce political scientists to invest in contributing to the domestic public debate by – also – writing in Dutch and for a wider public. We shall see: the next assessment of research quality in political science is scheduled for 2019.

References

Andeweg, R.B. and B. Vis (Eds) (2015), *Van Politisering naar Professionalisering; Politicologie in Nederland*, Oudewater: NKWP, 126 pp.

Bensing, J. , R. Andeweg, Ph. Franses, B. Meyer, C. Prins, K. Schuyt (2013), *Towards a Framework for the Quality Assessment of Social Science Research*, Amsterdam: Royal Netherlands Academy of Arts and Sciences, 43 pp.

Standard Evaluation Protocol 2015-2021; Protocol for Research Assessments in the Netherlands (2014) Association of Universities in the Netherlands (VSNU), Royal Netherlands Academy of Arts and Sciences (KNAW), Netherlands Organization for Scientific Research (NWO), 32 pp.

Van Drooge, L., S. de Jong, M. Faber and D. Westerheijden (2013), *Twintig Jaar Onderzoeksevaluatie; feiten & cijfers*, Rathenau Instituut (www.rathenau.nl), 19 pp.

Verdun, A., D. Farrell, O. Gabriel, C. Hay, H. Heinelt, K.E. Jørgensen, M. Kenny (2014), *Research Review Political Science 2007-2012*, Utrecht: Qanu, 69 pp.